IJCSN International Journal of Computer Science and Network, Volume 3, Issue 4, August 2014
ISSN (Online) : 2277-5420 **www.IJCSN.org**
Impact Factor: 0.274

Applying Power Law on Texture Structure to Identify the Writing Style in Ancient Manuscripts

[1] Ahmad Abd Al-Aziz, [2] Mervat Gheith

[1] Business Information Systems, Canadian International College (CIC)
Cairo, Egypt

[2] Computer Science and Information, Institute of Statistical Studies and Researches (ISSR) Cairo University
Cairo, Egypt

Abstract - This work presents the preliminary results for identifying the ancient handwritten writing styles in Arabic manuscripts. The aim of this work is to discriminate between different writing styles appear in different dated ancient Arabic manuscripts in three Islamic historical ages (Contemporary, Ottoman and Mamluk) with objective to evaluate the capability of applying Zipf law on texture structures extracted by Spatial Gray Level dependency (SGLD) method, our approach relies on applying SGLD on handwritten ancient text image to extract the texture structures then applying Zipf's distribution on document image texture structure. Based on this method we extract a feature presents the document image structure distribution and evaluate its efficiency for writing style identification.

Keywords - *Writing Style Identification; Spatial Gray Level Dependency; Power Law; Texture Analysis.*

1. Introduction

One of the main goals of a paleographer and historian experts is to locate handwritten materials. Consequently, a very crucial problem in paleographic field of study is to identify the author of a given manuscript [1]. Automatic recognition of handwriting styles is one of problems in ancient text document analysis due to large variation of writing styles in a population [2]. The variation of writing styles is not the only challenge image processing techniques can solve, but there are a lot of difficult things, for examples: (1) recognizing character fonts, (2) separating cursive characters into separate characters, and (3) distinguishing characters that have the same shape but have different meaning such as the character "o" and number "0" in English language and "ا" (Alef as character" and "١ " (One as digit) in Arabic language [3]. In addition to previous challenges, writing styles detection faces other problems occur in ancient manuscripts as deteriorated letters and some information as the author may be lost. [4]. The concept behind the writer recognition system is automatically or semi-automatically recognizing who is writing given certain manuscript, by applying two different tasks: Writer identification and writer verification. Writer identification determines which writer provides a given handwriting. Writer verification aims to decide on two handwritten manuscripts and determine if they are written by the same writer or by two different writers [5]. The benefit of automatically identifying the writing styles in ancient manuscripts is it can help paleographers and historian experts to analyze the manuscripts by checking for example whether a given part of manuscript image is contained in a stored large database of manuscript images [6].

Writer recognition approaches can be categorized into two distinct families: (1) text-dependent approaches and (2) text-independent approaches: In text-dependent approaches, the writer must write exactly a predefined or a given text. The text-independent (global) writer recognition is a process of identifying or verifying the identity of the writer without constraint on the text content [5]. The global technique of writing style detection gives more information about who the author is than the text contents itself and it enables to show if a document is an original or a copy [4].

This approach should guarantee the independency out of the text content, the writer's personal style, the language used and letters frequencies. The methods of global technique fall into two major categories: (1) Statistical and (2) Structural; *Statistical category* characterizes texture by the statistical distribution of the image intensity **[7]**; *Structural category* describes texture by identifying structural primitives and their placement rules. They are suitable for textures where their spatial sizes can be described using a large variety of properties. One of the most widely used methods is co-occurrence evaluated from Spatial Gray-Level Dependence SGLD which is a join probability to observe the same intensity value between two different pixels according to their spatial relation. SGLD matrix has a set of characteristics, it is

identical on different text areas of the same document and is robust to noise and does not require any image segmentation or layout analysis [8], The SGLD is not only identical on different text areas of the same document but also similar to all documents in entire manuscript of the same writer [9]. In the linguistic analysis field of study, Zipf's law used efficiently to represents the distribution of frequencies of words in the form of power law discovered by Zipf in 1949 in written English texts [10].

In the field of digital images processing, Zipf's law has been used in [11] to evaluate the distortion of compressed images. In [12] the power law used efficiently for artificial objects detection in natural environments by using a method based on Zipf's law, and this method has advantage of totally independent from the object shape. In [13] a detection of the existence of hidden message in LSB steganography is done, the detection theory is based on statistical analysis of pixel patterns using a Zipfness measure between successive bit planes [13].

The main objective of this work is to evaluate the applying of Zipf's law on the texture structure extracted from SGLD method to discriminate between different writing styles in ancient Arabic manuscripts. This work differs from other works in which Zipf's law applied on document image texture structures instead of quantize the original based on individual pixel values. This work organized as follows: in section II and III, we describe the theoretical background of SGLD and Zipf's law respectively. Related work described in section IV. Our methodology described in details in section V. in section VI we present the experimental results and discussion and finally in section VII we conclude the main points in this work and list some important points related to future work.

2. Spatial Gray Level Dependence (SGLD)

The co-occurrence can be evaluated from the SGLD which is a join probability to observe the same intensity value between two different pixels according to their spatial relation. [8]. SGLD also known as spatially dependencies matrix and has been known as a powerful method to represent the image textures which can be described as patterns of "non-uniform spatial distribution" of grey scale pixel intensities [14]. Mathematically, a co-occurrence matrix C is defined over an N × M image I, parameterized by an offset (Δx, Δy) [15] as:

$$C_{\Delta x,\Delta y}(i,j)=\sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases}1, & \text{if } I(p,q)=i \text{ and } I(p+\Delta x,q+\Delta y)=j\\ 0, & \text{otherwise}\end{cases} \quad (1)$$

manuscript document images are shown in figure 1.

Fig.1. Spatial graylevel ependence matrices representations of different writer styles in different Historical ages

From each generated matrix, 14 statistical measures are extracted including: angular second moment, contrast, correlation, variance, inverse different moment, sum average, sum variance, sum entropy, difference variance, difference entropy, information measure of correlation I, information measure of correlation II, and maximal correlation coefficient. The measurements average the feature values in all four directions 0^o, 45^o, 90^o and 135^o [16].

3. Spatial Gray Level Dependence (SGLD)

Zipf law is an empirical law and relies on a power law [17]. The main concept of this law states that in phenomena figured by a set of topologically organized symbols, the distribution of the occurrence numbers of n-tuples named patterns is organized in such a way that the apparition frequency of the patterns M_1, M_2, …. M_n, noted N_1, N_2,..,N_n, are in relation with rank of these pattern when sorted with respect to their occurrence frequencies. The following relation holds:

$$N_{\sigma(i)} = K \times i^{a} \quad (2)$$

$N_{\sigma(i)}$ represents the occurrence number of the pattern with rank i. k and a are constants. The power law is characterized by the value of the exponent a. k is more linked to the length of the symbol sequence studied. The relation is not linear but a simple transform leads to a linear relation between the logarithm of N and the logarithm of the rank. Then, the value of exponent a can be easily estimated by the leading coefficient of the regression line approximating the experimental points of the 2D graph called Zipf graph ($log_{10}(i)$, $log_{10}(N\sigma(i))$) with $i=1$ *to* n.

One way to achieve the approximation is to use the least square method. As points are not regularly spaced, the points of the 2D Zipf graph are re-scaled along the horizontal axis. The validity of this law has been observed in many domains but rather for mono dimensional signals [18].

4. Related Work

The SGLD used efficiently in ancient document recognition and classification. In [19] they proposed the use of multichannel Gabor filtering and SGLD matrices to characterize the writing style of writers in ancient manuscripts. For the global features, a study has been done on an implementation in [20] which was based on Haralick's features. Although they claim that SGLD gives very significant result, it required the testing and training images to be in the same dimension due to the multiplication of matrices in SGLD [21]. SGLD used in writer recognition in [22], they applied the SGLD to extract several features to characterize the writing style of ancient Latin and Arabic manuscripts of the middle-ages [8].

Power law has different applications on image such as detecting region of interest in image as in [10]. They proposed that Zipf law and inverse Zipf law can be used to detect regions of interest in an image. In their work they quantize the grey level by dividing the grey scale into a small number of classes and assign at each pixel the value of the class. The main disadvantage of previous coding is that the number of symbols used to represent the grey levels is limited, so if there are two images with different patterns the coded pattern that is generated it may be the same in both images. Another application applied Zipf law on image processing proposed in [4] for writer identification, they used K-mean algorithm for coding the image and applied Zipf law to identify the writer in document image, although the result of this work was not good enough but it encouraging us to test the capability of applying Zipf law on SGLD texture structure matrix in order to overcome the disadvantage of image quantization used method.

5. Methodology

In this work we start by acquiring the old Arabic documents image, converting it into gray level scale to prepare it for applying SGLD method, each cell in the SGLD matrix contains the frequency N of occurrence of each pattern (co-occurrence of pixel with its relationship with other pixel) in the original image according to offset (Δx, Δy) parameter. To extract features, SGLD matrix should be normalized in order to calculate the descriptive statistical features (mean, standard deviation, contrast, homogeneity…etc.), but in our method we decide to keep the frequencies of each pattern without normalization in order to apply the zip's law. Patterns in SGLD matrix are sorted in the decreasing order of their appearance frequency *N*, and the frequency for each pattern is plotted with respect to its rank *i* (the most frequent pattern has rank =1) in a double-logarithmic scale diagram and ignoring pattern rank *i* with frequency 0. In this work we calculate the area under Zipf's curve as a feature selection to discriminate between different writing styles, the block diagram of our methodology and example of applying the Zipf's distribution by using the output of SGLD matrix are shown in figure 2 and figure 3 respectively.

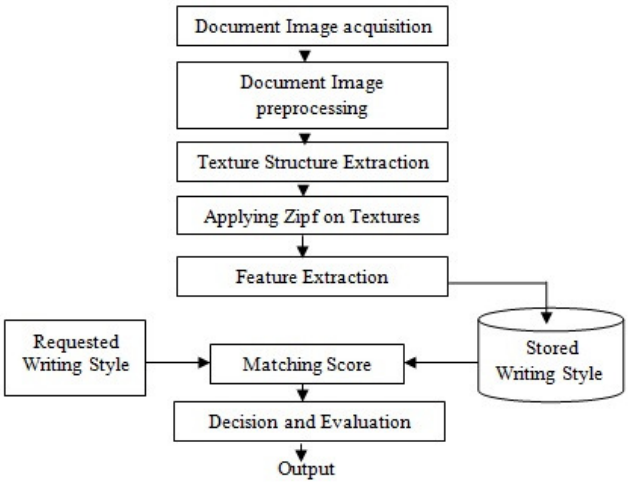
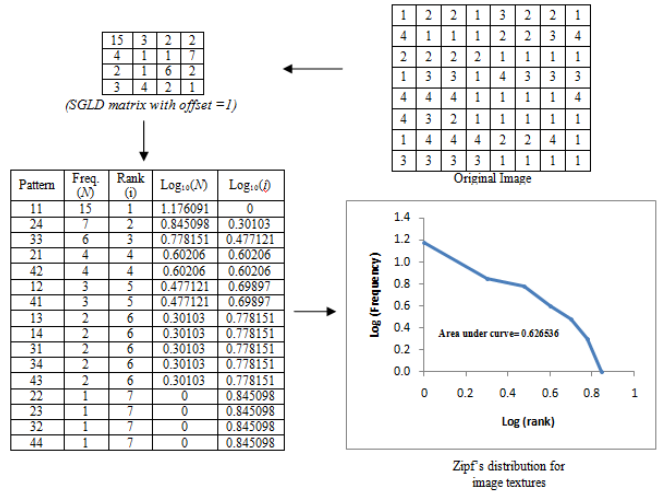


Fig.2. Block diagram of our methodology



Pattern	Freq. (N)	Rank (i)	$Log_{10}(N)$	$Log_{10}(i)$
11	15	1	1.176091	0
24	7	2	0.845098	0.30103
33	6	3	0.778151	0.477121
21	4	4	0.60206	0.60206
42	4	4	0.60206	0.60206
12	3	5	0.477121	0.69897
41	3	5	0.477121	0.69897
13	2	6	0.30103	0.778151
14	2	6	0.30103	0.778151
31	2	6	0.30103	0.778151
34	2	6	0.30103	0.778151
43	2	6	0.30103	0.778151
22	1	7	0	0.845098
23	1	7	0	0.845098
32	1	7	0	0.845098
44	1	7	0	0.845098

Fig.3. Applying zip's law on SGLD matrix

We proposed that our method can discriminate between different writing styles. In this work we applied our methodology on three different writing styles in different document images to evaluate its ability to discriminate between them as shown in figure 4.

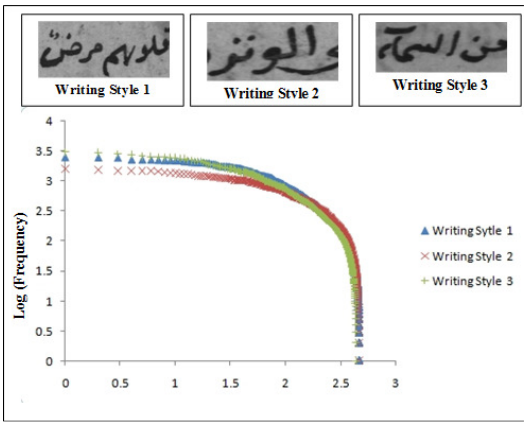


Fig.4. Three different zipf's distribution for three different writing styles

In figure 3 we showed that by applying Zipf's distribution on SGLD matrix's output for different writing styles, each writing style has different Zipf's distribution. In our methodology we apply Hamming distance "(3,)" as one of similarity measurements between the requested and stored writing style in order to identify the most similar writing style for the requested writing style.

$$Dist\ (I, I') = \sum_{i=1}^{d} |S_i - S'_i| \qquad (3)$$

6. Experimental Results and Discussion

The dataset includes 100 document images in three different historical ages; Contemporary [1220 Hijri till present], Ottoman [923 Hijri : 1220 Hijri] and Mamluk [648 Hijri : 923 Hijri] in 30 different writing styles of 30 author and copyist collected from 30 books from the "*Dar Al-Kotob Al-Masri*a" library in Egypt.

In order to evaluate the performance of applying Zipf law on texture structure generated by SGLD matrix, we calculate the area under Zipf's curve for each writing style in dataset and compared the tested writing style by calculating the Hamming distance between the requested query document image and stored dataset. Our method has been tested by 30 different writing styles for 30 different authors and copyists in Contemporary, Ottoman and Mamluk ages. The results show that 91.6 % recognition accuracy achieved in contemporary age writing style samples, 89.2% recognition accuracy achieved in Ottoman age writing style samples and finally 84.2% recognition accuracy achieved in Mamluk age as shown in figure 4, some parts of tested document is shown in figure 5 . These results show that applying Zipf law on texture structure generated from SGLD is efficient in identifying different writing styles in historical Arabic manuscripts.

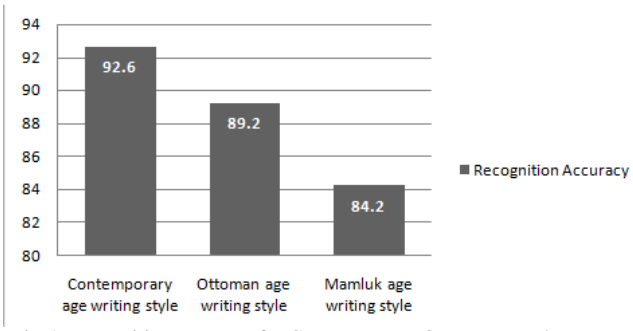


Fig.5. Recognition accuracy for Contemporary, Ottoman and Mamluk age's writing styles

Fig.6. Part from the experiment results

7. Conclusion and Future Work

This work has been done with objective to discriminate between different writing styles in different Islamic historical ages: Contemporary, Ottoman and Mamluk. We aimed to extend the approach of applying Zipf''s law in image applications by proposed a new approach based on texture structures generated by SGLM method to identify the writing styles in ancient Arabic manuscripts. This method has an advantage of analyzing the structure of document image patterns instead of image quantization limitations. By applying Zipf's law on generated SGLD we compute the area under Zipf curve for each writing style and evaluate the experimental results. These results show that applying this method is efficient in identifying different writing styles in different historical Arabic manuscripts. The main disadvantage is that the processing time of SGLD is high in addition the query and stored writing style image should be processed in the same orientation in SGLD matrix generation. In our future work we plan to add more features e.g. curve slope with area under curve which is evaluated in this work in order to increase the efficiency of writing style identification task.

References

[1] F. Aiolli and M.Giollo, "A Study on the writer identification task for paleographic document analysis", Proceedings of the 11th IASTED International conference on artificial intelligence and applications (AIA), 2011.

[2] L.Schomaker, G. Abbink and S. Selen, "Writer and writing-style classification in the recognition of online handwriting", in: Proceedings of the European workshop on handwriting analysis and recognition: A European perspective, pp. 12-13. 1994.

[3] E. Vellingiriraj and P. Balasubramanie, "Recognition of ancient tamil handwritten Characters in palm manuscripts using genetic algorithm. International journal of scientific engineering and technology,vol. 2, pp. 342-346, 2013.

[4] R. Pareti and N. Vincent, "Global method based on pattern occurrences for writer identification", 10th International workshop on frontiers in handwriting recognition, 2006.

[5] C. Djeddi, L. Souici-Meslati and , A. Ennaji, "Writer recognition on arabic handwritten documents", in: Proceedings of the 5th international conference on Image and Signal Processing, pp. 493-501, 2012.

[6] R. Herzog, A. Solth and B. Neumann, "Using Harris Corners for the retrieval of graphs in historical manuscripts", 12th International conference on document analysis and recognition, 2013.

[7] J. Brenard, Digital Image Processing, Springer-Verlag Berlin, Germany, 2005.

[8] V. Eglin, F. Lebourgeois, S. Bres, H. Emptoz, Y. Leydier, I. Moalla, and F. Drira, "Computer assistance for digital libraries: contributions to middle-ages and authors manuscripts exploitation and enrichment". 2nd International conference on document image analysis for libraries (DIAL'06), pp. 265-280, 2006.

[9] A. Al-Aziz, M. Gheith and A. Sayed, "Recognition for old Arabic manuscripts using spatial gray level dependence (SGLD)", Egyptian Informatics Journal, vol. 12, pp. 37–43, 2011.

[10] Y. Caron, P. Makris and N. Vincent, "Use of power law models in detecting region of interest", Pattern recognition, vol. 40, pp. 2521 – 2529, 2007.

[11] N. Vincent, P. Makris and J. Brodier, "Compressed image quality and Zipf's law, in: Proceedings of international conference on signal processing (ICSP-IFIC-IAPRWCC2000), pp. 1077–1084, 2000.

[12] Y. Caron, P. Makris, N. Vincent, "A method for detecting artificial objects in natural environments", in: Proceedings of international conference on pattern recognition (ICPR2002-IAPR), pp. 600–603, 2002.

[13] L. Lakhdar and H. Merouani, "A novel technique of steganalysis in uncompressed image through zipf's law", International journal of computer applications, vol. 40, 2012.

[14] M. Roomi and S. Saranya, "Bayesian classification of fabrics using binary co-occurrence matrix. International journal of information sciences and techniques (IJIST) ,vol. 2, 2012.

[15] W. Boussellaa, H. El-Abed and A. Zahour, "A concept for the separation of foreground/background in arabic historical manuscripts using hybrid methods", 7th International symposium on virtual reality, archaeology and cultural heritage VAST, 2006.

[16] M. Sharma, M. Markou, and S. Singh, "Evaluation of texture methods for image analysis", in: Proceedings of the 7th Australian and New Zealand intelligent information systems conference, pp. 117–121, 2001.

[17] J. Eakins, "Content base image retrieval – can we make it deliver?", 2nd UK Conference on image retrieval, 1999.

[18] K. Melessanaki, V. Papadakis, C. Balas and D. Anglos, "Laser induced breakdown spectroscopy and hyper-spectral imaging analysis of pigments on an illuminated manuscript", Spectrochimica Acta Part B 56, pp. 2337-2346, 2001.

[19] L.M. Al-Zoubeidy, and H. F. Al-Najar, "Arabic writer identification for handwriting images". In: Procceeding of International Arab Conference on Information Technology, pp. 111-117, 2005.

[20] I. Moalla, M. Alimi, F. Lebourgeois, and H. Emptoz, "Image analysis for palaeography inspection, 2nd International conference on document image analysis for libraries (DIAL'06), pp. 303–311, 2006.

[21] M. S.Azmi, K. Omar, M. F. Nasrudin, A. K. Muda, A. Abdullah and K. Ghazali, "Features extraction of arabic calligraphy using extended triangle model for digital jawi paleography analysis", International journal of computer information systems and industrial management applications, vol. 5, pp. 696-703, 2013.

[22] H. Said, T. Tan and K. Baker, "Personal identification based on handwriting", Journal of pattern recognition society, Vol. 33, pp. 149-160, 2000.

.

Ahmad Abd Al-Aziz received his M.SC., in computer science from Institute of Statistical Studies and Researches (ISSR), Cairo University, he enrolled in PhD in computer science in (ISSR), Cairo University. He joined Canadian International College (CIC), in 2007 as assistant lecturer. His main areas of research interest are Pattern Recognition, Social Network analysis, Big data, Natural Language processing, sentiment analysis. He had multiple disciplines background: social sciences (since he gained his PhD in behavior sciences from Ain Shams University) and computer science empowered him to scale out his research areas in several topics related to applying information theories in social sciences. Ahmad published two scientific papers in image analysis and pattern recognition.

Mervat Gheith is assistant professor at Institute of Statistical Studies and Researches (ISSR), Cairo University, her main research interests are artificial intelligence, pattern recognition, image processing and natural language processing. Mervat published several scientific papers in fields of natural language processing and artificial intelligence.